Technical Report 1272

Scoring Situational Judgment Tests Using Profile Similarity Metrics

Peter J. Legree, Robert Kilcullen and Joe Psotka
U.S. Army Research Institute

Dan Putka
Human Resources Research Organization

Ryan N. Ginter
George Mason University
Consortium Research Fellows Program

July 2010

United States Army Research Institute
for the Behavioral and Social Sciences

U.S. Army Research Institute
for the Behavioral and Social Sciences

Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

MICHELLE SAMS, Ph.D.
Director

Technical Review by

Christopher Vowels, U.S. Army Research Institute
Brian Tate, U.S. Army Research Institute

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
September 2010	Final	August 2009 – June 2010

4. TITLE AND SUBTITLE	5a. CONTRACT/GRANT NUMBER
Scoring Situational Judgment Tests Using Profile Similarity Metrics	
	5b. PROGRAM ELEMENT NUMBER 622785
6. AUTHOR(S)	5c. PROJECT NUMBER A790
Peter J. Legree, Robert Kilcullen and Joe Psotka (U.S. Army Research Institute for the Behavioral and Social Sciences); Dan Putka (Human Resources Research Organization); and Ryan Ginter (George Mason University)	5d. TASK NUMBER 270
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: DAPE-ARI-RS 2511 Jefferson Davis Highway Arlington, VA 22202-3926	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 2511 Jefferson Davis Highway Arlington, VA 22202-3926	10. SPONSOR/MONITOR'S ACRONYM(S) ARI
	11. SPONSORING/MONITORING Technical Report 1272

12. DISTRIBUTION AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Subject Matter POC: Peter J. Legree

14. ABSTRACT

This paper describes the application of profile similarity metrics to score Situational Judgment Tests (SJTs) that utilize rating scales to register examinee responses. The paper presents and discusses mathematical analyses that decompose distance-based measures into component indices based on correlation, dispersion and elevation metrics. The mathematical analyses demonstrate that distance measures represent a mixture of variance that can be associated with these separate components. Comparing the validities of distance and component indices using Leader Knowledge Test (LKT) data supports conclusions that the use of component indices (i.e., correlation, dispersion and elevation scores) improves the validity of SJTs that utilize rating scales.

15. SUBJECT TERMS

Situational Judgment Tests, Profile Similarity Metrics, Leader Knowledge Test, junior officers, lieutenants, captains,

SECURITY CLASSIFICATION OF:			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21.RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	**Unclassified**	**Unclassified**	Unlimited	35	Ellen Kinzer Technical Publication Specialist 703-602-8049

Standard Form 298

Technical Report 1272

Scoring Situational Judgment Tests Using Profile Similarity Metrics

Peter J. Legree, Robert Kilcullen, and Joseph Psotka
U.S. Army Research Institute

Dan Putka
Human Resources Research Organization

Ryan N. Ginter
George Mason University
Consortium Research Fellows Program

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

September 2010

Army Project Number
622785A790

Personnel, Performance and Training Technology

SCORING SITUATIONAL JUDGMENT TESTS USING PROFILE SIMILARITY METRICS

EXECUTIVE SUMMARY

Research Requirement:

Situational judgment tests (SJTs) present examinees with scenarios involving a dilemma, problem, or conflict and ask them to select the most appropriate response to the scenario from several options or to rate the appropriateness of each option. With the latter method, SJT items (represented by each alternative) are typically scored according to the mean ratings of experts on each response alternative/item (Motowidlo, Dunnette & Carter, 1990). Agreement between expert and examinee ratings is then assessed in terms of distance scores (e.g., Sternberg et al., 2000; Moros, 2008; Grim, 2010; Wagner & Sternberg, 1985). The use of distance scores has been increasingly preferred to other methods because SJTs can be time-intensive to administer and this method of scoring SJTs is very efficient.

A drawback associated with distance scores is that they confound variance associated with the way individuals use the rating scales. For example, poor distance scores will be computed for respondents who center their ratings in the upper half of the scale, even if their ratings are highly correlated with the scoring key. This occurs because the examinee's ratings will diverge sharply from those in the scoring key for those items that are keyed in the lower half of the scale. Likewise, respondents whose ratings vary too much, or too little, will receive poor distance scores even if their ratings are highly correlated with the scoring key. Previous SJT analyses have not systematically investigated relationships among measures of absolute agreement (distance), association (correlation), centering (elevation differences between the key and respondent rating profiles), and dispersion (standard deviation or variance). T present research will examine the relationships among these measures in order to shed light on advantages and limitations of using similarity metrics, like distance scores, to score SJTs..

Procedure:

Mathematical analyses were conducted by conceptualizing SJT scoring keys and examinee rating sets (i.e., response profiles) as vectors with profile similarity metrics proposed to assess individual differences in the quality of examinee ratings. Distance formulae were expanded to identify relationships among indices of association, dispersion and elevation.

Statistical analyses, which were based on implications from the mathematical analyses, were conducted to understand and improve the validity of the Leader Knowledge Test (LKT) as an SJT.

Findings:

The formula for the mean distance squared between the *n* elements in an examinee response profile, **X**, and the scoring profile, **K**,

$$D^2 = \sum(X_i\text{-}K_i)^2/n \text{ for } i = 1 \text{ to } n,$$

mathematically expands to:

$$D^2 = (X_{mean} - K_{mean})^2 + ((n\text{-}1)(sd_x^2 + sd_k^2 - 2sd_x sd_k C))/n.$$

Where X_{mean} is the mean value in vector **X;** K_{mean} is the mean value in the vector **K;** sd_x is the standard deviation of values in vector **X**; sd_k is the standard deviation of values in vector **K**; and *C* is the product moment correlation between values in **X** and **K**.

We compared the overall validity of LKT C-Scores (based on the correlations between the scoring key, **K**, and each respondent vector, **X**, computed separately for the two LKT subscales) against the corresponding D^2-scores (based on the mean squared distance between elements in the scoring key, **K,** and each respondent vector, **X**). We also evaluated indices of dispersion (based on sd_x) and elevation (based on X_{mean}). Analyses conducted separately for the two LKT subscales show:

- LKT C-score correlations with rank of the Soldiers (r = .55 & .67) were substantially greater than the corresponding LKT D^2-score correlations (r = .41 & .36).

- LKT C-score and Elevation scores correlated with personality indices that relate to leadership (r ranged from up to .45).

Utilization and Dissemination of Findings:

Analyses show that distance-based scoring algorithms may minimize validity estimates by confounding correlation, elevation, and dispersion effects. Use of correlation-based scores were shown to improve understandings regarding the construct validity of the LKT. These analyses confirm that rating patterns on carefully constructed scales reflect individual differences in expertise. These results provide a basis for the creation and scoring of valid SJTs that efficiently utilize administration time.

SCORING SITUATIONAL JUDGMENT TESTS USING PROFILE SIMILARITY METRICS

CONTENTS

LIST OF TABLES

LIST OF FIGURES

SCORING SITUATIONAL JUDGMENT TESTS USING PROFILE SIMILARITY METRICS

INTRODUCTION

Situational Judgment Tests (SJTs) present open-ended scenarios to examinees along with descriptions of alternative ways of addressing each scenario. The scenarios often summarize interpersonal problems or situations, and the alternatives describe realistic responses to those problems and situations. After reading the scenarios and alternatives, examinees are asked to evaluate the proposed alternatives. In the most straightforward response format, examinees are requested to identify the most appropriate or the best alternative for each scenario, much like a multiple-choice test item. SJT scoring keys are usually developed by surveying experts to quantify the correctness or appropriateness of each of the alternatives. Expert keying approaches are required for SJTs because the appropriateness of response-alternatives is inherently ambiguous, and it is usually impossible to identify correct responses based on theory or doctrine (Motowidlo, Dunnette & Carter, 1990; McDaniel & Nguyen, 2001).

An important limitation with the SJT multiple-choice format is that the scales consume much reading time and may require over an hour to administer for test scores to be reasonably reliable. In order to reduce SJT reading requirements while maintaining scale psychometrics, researchers have increasingly asked respondents to rate the quality of all the alternatives for each scenario. The use of ratings allows more data to be collected per scenario so that the number of scenarios that must be included in the test can be substantially decreased. For example, a 50-scenario SJT that uses a multiple choice (5-option) format will yield 50 observations, while a 10-scenario SJT with 5-options per scenario that uses the ratings approach will yield 50 observations, thereby decreasing reading requirements by about 80%. With SJTs administered in this way, as opposed to asking examinees to select a single most appropriate response to a scenario, examinee scores can be computed as distances between examinee ratings and the mean score of those provided by experts (e.g., Sternberg et al., 2000; Moros, 2008; Grim, 2010; Weis, 2008).

While the ratings format clearly provides much more data per scenario than the single-choice format, the SJT literature has not considered implications associated with the use of the distance metrics used to score ratings-format SJTs from a mathematical perspective. This report identifies mathematical implications regarding the use of distance scores and, in so doing, offers suggestions for improving the validity and utility of these types of SJTs.

PROFILE SIMILARITY METRICS

When ratings data are collected for SJT scenarios, an examinee response profile can be described as a rating vector, **X**, with *n* elements where *n* equals the total number of rated alternatives. Likewise, the scoring rubric can be described as a scoring vector, **K,** also with *n* elements. Individual scores can then be computed by quantifying the correspondence between the scoring vector and the rating vector for each individual. In SJT applications that have used examinee ratings, the correspondence between a respondent's rating profile and the scoring key

is often represented by a measure of mean item distance, $D = \sum|X_i\text{-}K_i|/n$ (e.g., Weis, 2008; Muros, 2008; Sternberg et al., 2000; Wagner & Sternberg, 1985).

By conceptualizing respondent rating sets as profiles, the relevance of profile similarity metrics becomes explicit. Accordingly, the correspondence between a respondent rating profile, **X**, and the scoring key profile, **K**, can be quantified with (a) *Overall Measures* that assess the level of absolute agreement between the respondent and scoring key profiles, or (b) *Component Measures* that assess the similarity of the shape, elevation and dispersion of the two profiles (cf. Cronbach & Gleser, 1953). Using this distinction, some of the metrics that can be used to evaluate performance on ratings-based SJTs are listed in Table 1.

Overall measures of agreement include distance (D) and distance-squared (D^2) metrics, as well as endorsement metrics that are based on values in the scoring profile (cf. Wagner & Sternberg, 1985; Moros, 2008; Grim, 2008; Mayer, Caruso & Salovey, 1999). Distance scores are computed as the mean (D) or mean square (D^2) of the absolute values of the differences between items in **X** and **K** so that superior performance is indicated by lower distance values. Endorsement Ratio measures are computed using values in **K** that quantify the mean proportion of experts endorsing each action to weight the values in each response profile, **X**, so that superior respondent profiles are indicated by higher values. While overall measures of agreement may vary computationally, they are conceptually similar because these measures are influenced by the similarity of the shape (correlation), elevation (mean rating) and variance of the respondent and scoring profiles.

Table 1. Situational Judgment Test Measures

Overall Measures	**Component Measures**
Mean Squared Distance: $D^2 = \sum(X_i\text{-}K_i)^2/n$	Mean Standardized Distance: $D_{stand} = \sum\|z_x - z_k\|/n$
Mean Distance : $D = \sum\|X_i\text{-}K_i\|/n$	Mean Squared Standard Distance: $D^2_{stand} = \sum(z_x - z_k)^2/n$
Endorsement Ratios	Correlation scores: $r = 1 - \sum(z_x - z_k)^2/2(n\text{-}1)$
Factor scores	Percent Correct
	Dispersion: sd_x
	Elevation Difference: Eldif = $\|X_{mean} - K_{mean}\|$

We use the term *Component Measure* to refer to indices based on the shape (correlation), elevation (mean rating) and dispersion of the respondent and scoring profiles. Measures of shape include standardized distance and correlation scores (i.e., *"C-scores"*). Standardized distance scores are computed after converting the elements in **K** and **X** to z-scores in order to control for individual differences in elevation and dispersion. It is important to note that correlation scores are formulaically similar to standardized distance scores. This similarity is apparent by comparing terms in conventional formulae for the mean standardized distance, $D_{standardized} = \sum|z_x - z_k|/n$, and the product moment correlation, $r_{x,k} = 1 - \sum(z_x - z_k)^2/2(n\text{-}1)$ (cf. Cohen, Cohen, West & Aiken, 2003, equation 2.2.4). In addition, Mean Squared Standardized Distance scores can be directly computed from the correlation estimates:

$D^2_{standardized} = \sum(z_x - z_k)^2/n = 2(1 - r_{x,k})(n\text{-}1)/n$. In other words, Mean Squared Standardized Distance scores are linearly related to C-scores and entirely redundant.

In addition to measuring shape, component measures can also be computed to assess individual differences in the elevation (rating mean) and dispersion (rating standard deviation) of the respondent profiles. Perhaps because response strategies can be easily developed to influence elevation and dispersion (e.g., by inflating ratings on profiles or decreasing rating variance), these metrics have not been commonly used to score judgment tests. Yet, response strategies used to alter elevation and dispersion affect overall measures of agreement such as the distance (D-score) and distance-squared (D^2-score) metrics. Regardless of their susceptibility to faking, measures based on elevation and dispersion could provide useful secondary measures at least for some applications. Therefore, a closer look at the underlying mathematics is warranted.

Figure 1 depicts hypothetical rating profiles (i.e., vectors) for three respondents who rated the importance of five items on a 9-point scale as well as the scoring key used to assess the quality of the ratings. The scores have been deliberately chosen to illustrate strengths and weaknesses of each measure. This figure illustrates the point that conflicting results can be obtained with measures of distance and correlation. Although profiles **A** and **B** are both highly correlated with the scoring profile **K** ($r_{A,K} = r_{B,K} = .98$), the distance scores for profiles **A** and **B** indicate that profile **A** ratings are much more similar to the scoring key than profile **B** ratings, ($D_{A,K}^2 = .16$ vs. $D_{B,K}^2 = 3.84$). In this example, profile **B** might correspond to optimistic ratings obtained from a knowledgeable individual. The illustration also shows that the distance score for profile **C** is superior to the D-score for profile **B** (i.e., lower, $D_{C,K}^2 = 2.84$ vs. $D_{B,K}^2 = 3.84$) despite **C**'s correlation with the scoring standard, **K**, being much lower than **B**'s correlation ($r_{A,K} = .43$ vs. $r_{B,K} = .98$).

Figure 1. Scoring and Rating Profiles for Three Respondents

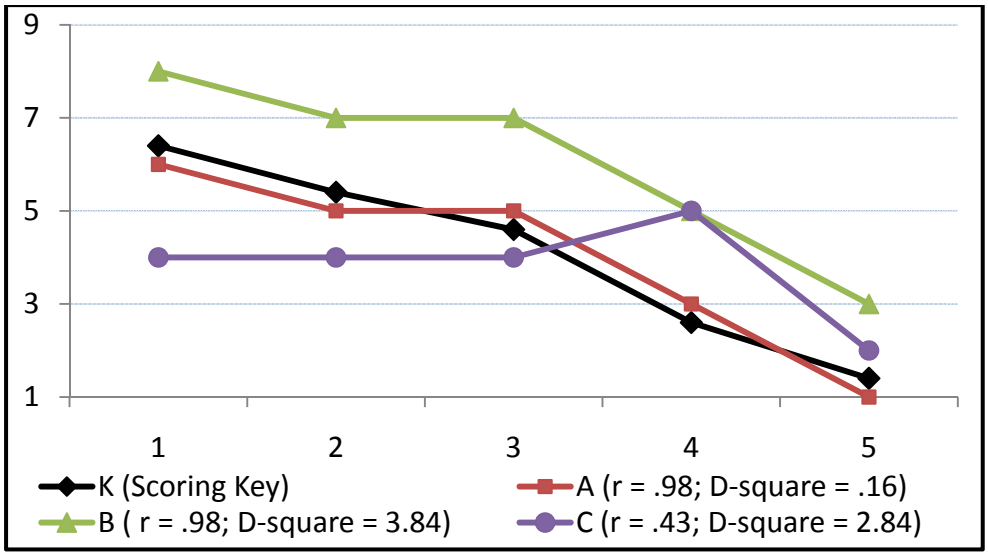


Figure 1 shows that (1) different strategies in the use of the rating scale by respondents may affect the extent to which distance scores accurately reflect individual differences in

performance on judgment tests and (2) highlights the importance of understanding relationships between measures of distance and correlation. It follows that use of these measures may lead to different interpretations. The following analyses explore the possibility that while distance measures utilize all of the available information in a rating profile, component measures based on the shape (correlation), elevation (mean rating) and dispersion (standard deviation) of the respondent and scoring profiles may not be properly weighted to optimize validity.

Derivations Based on the Mean Square Distance (D^2) Metric

To clarify the mathematical basis of the conflicts revealed in Figure 1, this next section analyzes distance and correlation based metrics from a mathematical perspective. Although D and D^2 metrics are mathematically related and tend to be highly correlated (e.g., the correlation between D and D^2 ranged from .98 to .99 in the LKT database described below), only the formula for D^2 can be expanded to explicate relationships among measures of absolute agreement and shape (i.e., distance and correlation). In the following equation, X_i and K_i correspond to observed ratings/values obtained from respondent vector, **X**, and the scoring key, **K**, for each item *i*. Equation 1 provides the standard formula for D^2.

$$D^2 = \sum(X_i\text{-}K_i)^2/n \text{ for item } i = 1 \text{ to } n. \quad (1)$$

By substituting $X_i = x_i + X_{mean}$ and $K_i = k_i + K_{mean}$, the effect of the difference in elevation between the respondent and scoring profiles ($\Delta_{Elevation} = X_{mean} - K_{mean}$) on D^2 can be isolated (x_i and k_i correspond to X_i and K_i centered, but not standardized.) Appendices A and B detail the following derivations. Substitutions yield:

$$D^2 = (X_{mean} - K_{mean})^2 + \sum(x_i\text{-}k_i)^2/n = \Delta^2_{Elevation} + \sum(x_i\text{-}k_i)^2/n. \quad (2)$$

Equation 2 can be used to understand the lack of consistency between the D-square estimates for profiles **A** and **B** (.16 versus 3.84) and their correlations with the scoring key, **K** (both .98). Although the deviation component is exactly the same, the elevation component is very different for the two profiles. Thus the difference between the D-square estimates is entirely due to differences in the means/elevations of profiles **A** and **B** with respect to the scoring key, **K**. In addition, Equation 2 can be manipulated to show that when a respondent provides no variance across his ratings (i.e., $sd_x = 0$ becasue all $x = 0$ after **X** is centered), D^2 is entirely determined by an elevation effect and the variance in the scoring profile, **K**:

$$D^2 = \Delta^2_{Elevation} + \sum k_i^2/n.$$

By expanding equation 2 and using statistical substitutions (i.e., $\sum x_i^2 = sd_x^2(n\text{-}1)$, $\sum k_i^2 = sd_k^2(n\text{-}1)$, $x_i = z_{xi}sd_x$, $k_i = z_{ki}sd_k$, and $\sum z_{xi}z_{ki} = r(n\text{-}1)$), interactions between the variance of x and k (i.e., sd_x^2 and sd_k^2) and the correlation (C) between the profiles on D^2 become clear. Substitutions yield:

$$D^2 = \Delta^2_{Elevation} + ((n\text{-}1)(sd_x^2 + sd_k^2 - 2sd_xsd_kC))/n. \quad (3)$$

Equation 3 shows that D^2 is equal to the sum of $\Delta^2_{Elevation}$, and a covariance function computed on sd_x, sd_k, n, and C. We use C instead of r to designate the correlation between the two profiles because we use "C-scores" to measure individual differences in the analyses that

follow. This equation demonstrates that D^2 mathematically combines components corresponding to elevation, variance and form (correlation). To the extent that these variables lack equipollence with respect to conceptually relevant criteria, it follows that separate metrics for these components may provide incremental validity in predicting relevant criteria.

Inspection of equation 3 also indicates that D^2 is minimized (made superior) by minimizing $\Delta^2_{Elevation}$ and maximizing C. In addition, when $\Delta^2_{Elevation}$ is constant and $C = 1$ (i.e., profiles **X** and **K** are perfectly correlated), D^2 will be minimized only when $sd_x = sd_k$.

For $C < 1$, the covariance term must be differentiated to identify the value of sd_x that minimizes the covariance term. Note that sd_k and n are constant for a specific application (i.e., across respondents). Setting:

$$\text{Covar} = ((n\text{-}1)(sd_x^2 + sd_k^2 - 2sd_x sd_k C))/n.$$

Then differentiating the covariance term with respect to sd_x, provides:

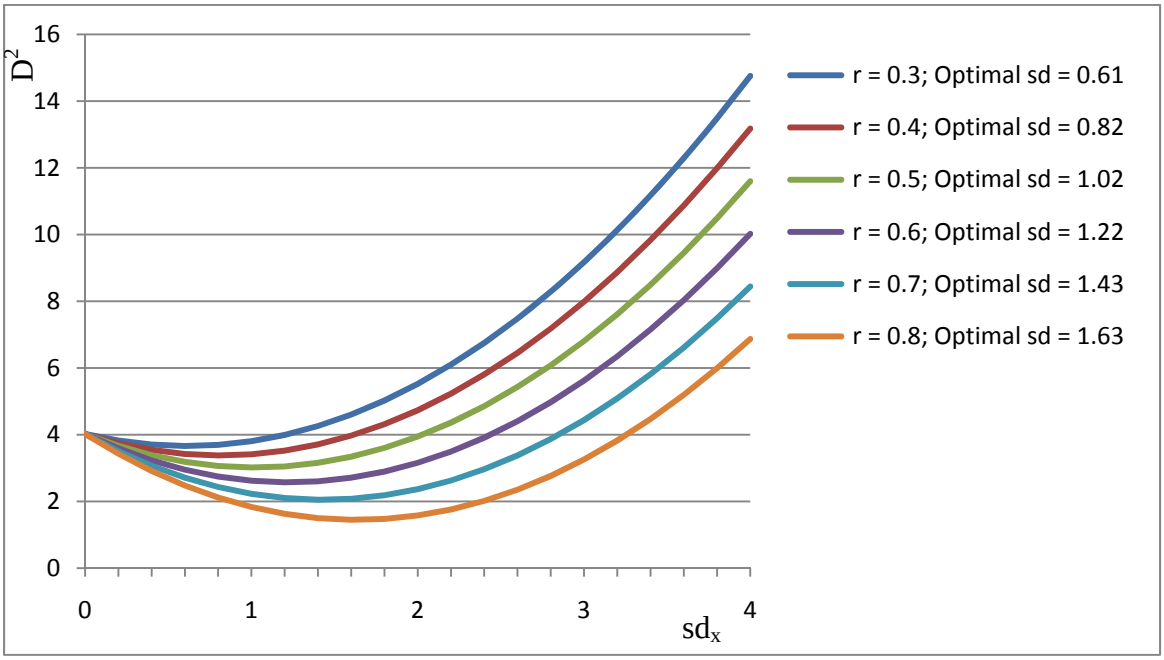
$$d(Covar)/d(sd_x) = (n\text{-}1)(2sd_x - 2sd_k C)/n.$$

The inflection point formulaically identifies the value of sd_x that minimizes the value of the Covar term for any value of C and sd_k. Solving for the inflection point, $d(Covar)/d(sd_x) = 0$, provides:

$$sd_x = sd_k C. \quad (4)$$

Equation 4 demonstrates that for $C < 1$, $sd_x = Csd_k$ will minimize the Covar term and its contribution to D^2. This equation shows that respondent D^2 scores are penalized for all $sd_x > sd_k$ and whenever $sd_x \neq sd_k C$. Figure 2 plots D^2 as a function of sd_x for specific C-scores (i.e., values of $r_{X,K}$). For $C < .50$, $sd_x = 0$ always provides a lower (i.e., superior) D^2 than $sd_x = sd_k$.

Figure 2. D^2 as a function of sd_x for specific C-scores with sd_k = 2.04



Scoring Key Effects on D^2

The importance of sd_x as a component of D^2 is increased when the scoring key profile is computed by averaging expert ratings for each item, $K_i = \sum E_i/p$ for i = 1 to p experts. This occurs because the standard deviation of the elements in the scoring key will be less than the mean standard deviation of the individual expert profiles, sd_{expert_mean}. Appendix B shows that sd_k is approximated by:

$$sd_k \approx (sd_e/p)(1 + \text{sqrt}(r_{mean})(p\text{-}1))/p \quad (5)$$

where r_{mean} equals the mean intercorrelation of expert ratings provided that individual expert rating profiles have similar variance.

Equation 5 shows that even experts would be required to limit the variance in their ratings to maximize the D^2 metric. Likewise, non-experts would generally need to provide ratings with much less variance than the average expert to maximize the D^2 metric. Equation 5 also suggests that less consistent groups of experts (e.g., journeymen versus technical experts) may be used to develop separate scoring keys that are highly correlated yet differ in variance. This implication is consistent with the rationale for consensus based assessment (Legree, Psotka, Tremble & Bourne, 2005).

Implications for Using Distance and Profile Similarity Metrics to Assess Judgment Test Ratings

When applied to scoring rating data obtained for judgment tests, Equation 3 indicates that individual differences in D^2 can reflect respondent differences in any of the components $\Delta^2_{Elevation}$, sd_x, and C. This occurs because K_{mean}, sd_k and n are specific to the judgment test

application and therefore are constant over respondents. While individual differences in *C* are relatively difficult to manipulate by respondent strategies (except perhaps to lower *C*), Equations 3 and 4 suggest that individual differences in D^2 could be manipulated by respondents by deliberately controlling $\Delta^2_{Elevation}$ and sd_x. In fact, whether intentional or not, such differences are readily seen in explorations of data (see below).

Although the goal of this analysis is not to develop SJT coaching strategies, Equation 3 suggests that respondents should center their ratings on the scale mean and minimize their variance to obtain superior distance scores. This follows because most SJTs have been constructed to have a broad range of response options, thus K_{mean} is often similar to the scale mean. This formula also provides the mathematical basis for understanding demonstrations that SJT distance scores can be improved by recoding extreme responses as more moderate (Cullen, Sacket & Lievens, 2006).

Strategy Implications of these Analyses

Equation 4 demonstrates that respondents who fully utilize the rating scales such that $sd_x \approx sd_k$, are penalized by D^2. Instead, for superior D^2-scores, sd_x should always be less than sd_k, and when *C* is low, sd_x should be much less than sd_k, i.e., $sd_k = C*sd_x$. Somewhat surprisingly, interpretation of the above equations shows that for $C < .50$, $sd_x = 0$ always provides a lower (i.e., superior) D^2-score than $sd_x = sd_k$. However, judgment test instructions frequently encourage respondents to utilize the entire rating scale. It follows that by following instructions respondents tend to be penalized by distance metrics such as D^2.

Collectively, these equations show that D^2 will be much more similar to a measure of the difference in elevation (e.g., $\Delta^2_{Elevation}$) or dispersion than to a measure of association under some conditions. We suggest that strategies, such as simply using the same rating for all questions, should not lead to superior measured performance; yet it can using D^2 measures. With C-score measures, such a strategy would be appropriately discarded.

Strategies, such as not using the full rating scale, or avoiding extreme ratings, should also not be rewarded if the judgment test is designed to reflect the respondents' knowledge. Again, C-scores appropriately assess such responding; whereas D^2 measures inappropriately rewards such behavior.

C-Scores as Superior Measures of SJT Performance

We suggest that C-scores should be conceptualized as the primary index of performance on SJTs that incorporate ratings data because distance measures are affected by elevation and dispersion in ways that interfere with accurate knowledge assessment. Furthermore, C-scores standardize the ratings and remove any elevation differences between **X** and **K**. Although elevation differences may be useful in some applications, generally SJTs are not designed to have directional elevation differences, so that respondent elevation based scores (e.g., $\Delta^2_{Elevation}$ or $Eldis^2$ scores) or variance based scores (Dispersion) are never computed as measures. If SJTs were ever designed to create a monotonic response dimension to assess ratings, it would change these interpretations drastically, but with the current state of knowledge, such a direction has not been taken. Under current conditions, C-scores should be the superior metric of choice.

Although this recommendation is contrary to current practices that routinely use D-scores or D^2-scores, we believe that few researchers would intentionally use SJT scoring algorithms that are affected by variance, elevation and form in deleterious ways. This practice would be tantamount to simply not scoring test protocols and instead using measures of the elevation and dispersion of respondent rating profiles to measure individual differences. Possibly, the use of D-scores has minimized interest in and understanding of correlation based scores because the simple distance formula, $D = \sum|X_i\text{-}K_i|/n$, cannot be mathematically converted to a form that explicitly contains a correlation component. This is why we expanded the D^2 formula, not the D formula.

Rather than simply ignoring the information associated with dispersion and elevation, these results actually suggest supplementing C-scores with these additional component metrics through regression techniques to fully utilize all the information available in SJT respondent rating profiles. In addition, scores based on individual rating means and variance may allow exploration of relationships between these additional metrics and conceptually related criteria. For example, hypotheses might explore relationships between confidence and knowledge based on assumptions of relationships between confidence and rating dispersion (cf. Chi, Glaser & Farr, 1988; Pleskac & Busemeyer, 2010).

These conclusions suggest that the following metrics may be useful in quantifying individual differences in respondent rating profiles on SJTs:

1. Correlation scores (C-scores) computed as the product moment correlation between each profile, **X**, and the scoring key, **K**. As described above, C-scores are mathematically equivalent to mean square standardized distance scores. C-scores = $r_{X,K}$.
2. Elevation distance scores ($Eldis^2$) computed as the squared difference in elevation between each respondent profile mean and the scoring profile mean. Therefore, $Eldis^2 = \Delta^2_{Elevation} = (X_{mean} - K_{mean})^2$.
3. Elevation scores computed as the mean respondent rating. Elevation = $\boldsymbol{X}_{mean} = \sum X/m$.
4. Dispersion scores computed as the standard deviation of each respondent ratings profile: Dispersion = sd_x.

APPLICATION

Overview

The above derivations mathematically show that measures based on distance-square (D^2) can be decomposed into separable component measures of shape (correlation), scatter (variance) and differences in elevation ($\Delta^2_{Elevation}$). Measures based on C-scores are identical to the D^2 measures, with elevation and variance differences removed, namely standardized D^2. While the mathematical derivations are independent of application and formulaically correct, the utility of the component measures is dependent upon application. For example, if the component measures were highly correlated with D^2 for a specific application, then the use of component measures would have little potential to improve the validity of SJTs beyond that obtained using a conventional distance measure. Therefore, one major goal of these analyses was to determine if and when additional component measures account for incremental variance in conceptually related variables in comparison to measures based on D^2. To assess the extent to which consensual scoring standards and profile similarity metrics may provide useful indices for SJT performance, we conducted analyses using the Leader Knowledge Test (LKT) database (described below). We do not assert that all conclusions based on the LKT database will generalize to all SJT databases, but we do offer these analyses as an initial foray into understanding the use of profile scoring metrics to improve the utility of SJTs.

One subtle limitation with the mathematical derivations is that SJT metrics have usually used measures of simple distance (D), while the distance-squared metric (D^2) was expanded in the derivations. To assuage any concerns that D and D^2-scores are only moderately correlated, we conducted preliminary analyses to estimate the correlation between the D and D^2–scores using the LKT database.

Based on consensus score theory (Legree, Psotka, Tremble & Bourne, 2005), we also expected that scoring standards based on respondent profiles would be highly correlated with standards based on expert responses but primarily differ in variance. These relationships are consistent with the derivation described above in the section, *Scoring Key Effects on D^2*. Therefore, we developed and compared scoring keys based on (a) the mean respondent rating for each item (Consensus Key), and (b) the mean expert rating for each item (Expert Key).

In summary, we conducted analyses to:

1. Assess the correlation between individual difference measures based on simple distance ($D = \sum|X_i\text{-}K_i|/n$) and squared difference ($D_{X,K}{}^2 = \sum(X_i\text{-}K_i)^2/n$) formulae.
2. Quantify the similarity between scoring keys based on expert and respondent means.
3. Quantify the similarity between scores based on expert and respondent scoring keys.
4. Document LKT scale psychometrics.
5. Quantify correlations and relationships among the following profile similarity metrics as individual difference measures:
    a. Mean squared item difference: $D_{X,K}{}^2$–scores = $\sum(X_i\text{-}K_i)^2/n$;
    b. Correlation: C-scores = $r_{X,K}$;
    c. Elevation-distance: Eldis^2-scores = $(X_{mean} - K_{mean})^2$;
    d. Elevation: Elevation-scores = X_{mean}; and

e. Dispersion-scores: SD_x.

6. Validate the LKT profile similarity metrics against measures of experience (rank) and personality (justified below) to determine if the refined measures provide more potent metrics of LKT performance than global agreement measures that are based on D^2.

Leader Knowledge Test (LKT)

The LKT was designed to assess knowledge of traits and skills that are relevant to leader performance in the U.S. Army. LKT development was based on expectations that leaders gain tacit knowledge regarding the importance of leader-relevant skills and traits through experience and reflections upon those experiences (Polanyi, 1966; Wagner & Sternberg, 1985; Sternberg et al., 2000). It follows that individuals assigned to leadership positions (i.e., higher ranks) should have more a refined understanding of the traits and skills that are required for effective leaders in the U.S. Army (Yukl, 2002). It is also believed that leader performance is positively associated with personality traits such as dependability, openness and agreeableness (cf. Yukl, 2002; Bartone, Snook, & Tremble, 2002; Bartone, Eid, Johnsen, Laberg, & Snook, 2009). This reasoning led to two general hypotheses regarding LKT performance:

H1: Rank correlates with LKT indices

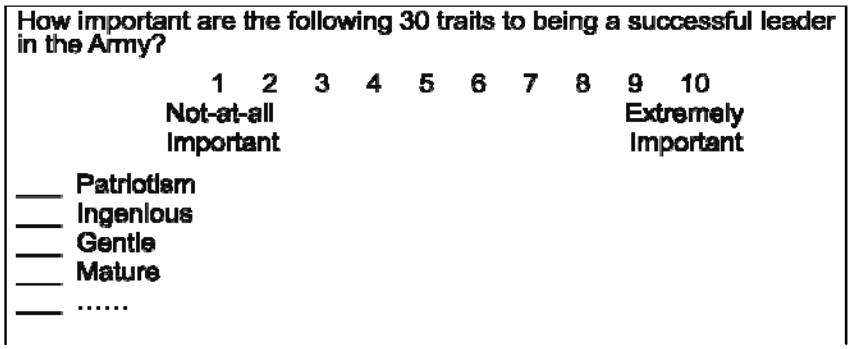
H2: Personality metrics correlate with LKT indices

While exploration of the LKT predictive validity is not possible with the current database, these hypotheses address the construct validity of the LKT. We utilized both measures of absolute profile agreement (i.e. D^2-scores) as well as related component measures of profile similarity (C-scores, Elevation, Dispersion, $Eldis^2$-scores) in our analyses. Although broad conclusions regarding the potency of profile scoring metrics require multiple databases, these results provide a baseline to guide expectations and analyses for similar databases.

LKT Construction

We surveyed the leadership literature (e.g., Yukl, 1994) to identify 15 characteristics and 15 skills that have been theoretically associated with effective leadership. In addition, we identified 15 characteristics and 15 skills that appear socially positive, and are used in the general job analysis literature, but have not been theoretically linked to effective leadership. The 30 characteristics and the 30 skills were assembled into two scales. Questionnaires instructions were designed to elicit respondent understandings of the relative importance of these characteristics and skills to military leadership. A 10-point Likert scale was incorporated into the judgment tests to enable individuals to come closer to a number matching psychophysical scale, which provides a continuous scale, and allows the respondent to register subtle differences in their understandings (Stevens, 1975). Figure 3 portrays example items from the LKT scale.

Figure 3. LKT Trait Scale Example Items.



Participants

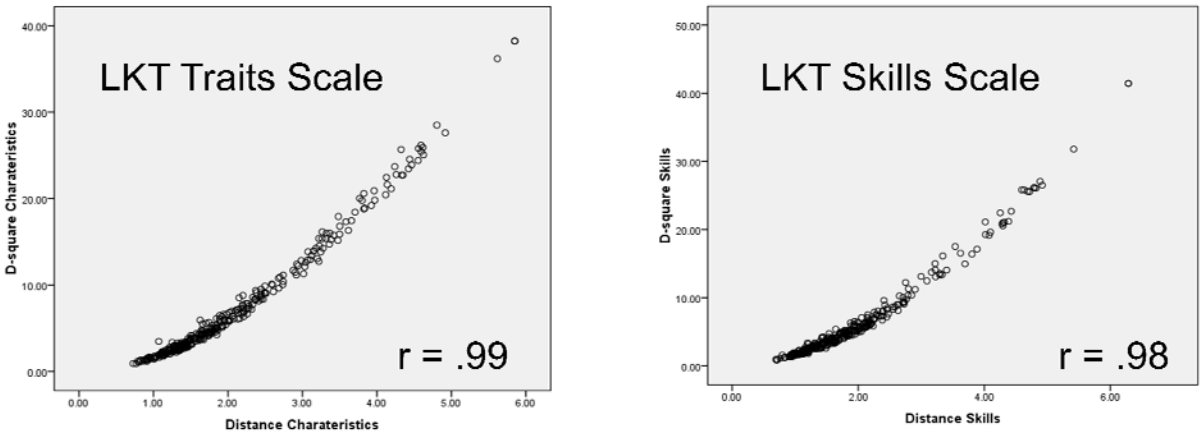
The sample consisted of 329 U.S. Army Soldiers including: 93 junior officers, 96 Non-Commissioned Officers (NCOs), and 124 enlisted Soldiers. These individuals were asked to complete the LKT scales. In addition, the NCOs and enlisted Soldiers completed the NEO Big-5 Mini-marker inventory and self-reported their rank. Project participation was voluntary.

To develop the expert scoring keys, 128 U.S. Army captains completed the LKT, and their ratings were averaged to create expert scoring keys for the LKT scales. The consensual scoring keys were computed by averaging the item ratings of the individuals in the validation sample. Analyses detailed below demonstrate high consistency between the two scoring keys. It follows that the expert key based on captain judgments is appropriate for scoring the LKT.

Results

Distance Scores. Because the mathematical derivations referenced D^2 and not D, while most SJT applications have used D, not D^2, the first set of analyses document the relationship between the two measures. Figure 4 summarizes the results and indicates a high degree of redundancy between the D and D^2–scores for both LKT scales, with both correlations at least .98. These scores were computed using the expert scoring key, similar results were observed using the consensus key (all $r \geq .97$). Based on these results, only analyses conducted using the D^2-scores are reported in the remaining tables, although similar results were consistently obtained for D-scores.

Figure 4. Relationship between D and D^2-scores for the LKT scales.



Scoring Key Analyses. Table 2 reports means, standard deviations and correlations for the Expert and Consensual Scoring Keys created for the two LKT scales. Results show that the values in both sets of scoring keys are highly correlated (r = .96) and have similar means yet differ in variance with the *SD* of the consensus key being less than the *SD* of the expert key. Table 2 also reports the correlations among D^2, $Eldis^2$, and C-scores computed using the Expert and Consensus Keys. Results show near-redundancy for individual difference measures (D^2, $Eldis^2$, and C-scores) computed using the Expert and Consensus scoring keys for both LKT scales, all r > .97.

Table 2. Correlations Between the LKT Expert and Consensus Keys (Key Values) and the LKT Scales based on those Keys (D^2-scores, C-scores, and Eldis2 Scores).

	K_{Mean}	K_{SD}	Correlations			
			Keys	D^2-scores	C-scores	$Eldis^2$
Trait Keys						
Expert	5.85	2.04	.96* (30)	.98** (328)	.98** (325)	.99** (329)
Consensus	5.59	1.26				
Skill Keys						
Expert	6.28	1.42	.96* (30)	.99** (315)	.97** (305)	.995** (315)
Consensus	6.09	0.75				

Note: n in parentheses; n = 30 for the correlations between the expert and consensus keys because there were 30 items in each key. * p < .01; ** p < .001.

Scale Psychometrics. The first three columns of Table 3 report the means, standard deviations and reliabilities for the D^2, C, Eldis^2, Dispersion and Elevation scores. Reliability coefficients were computed as coefficient alpha for the D^2 and C-scores, and were split-half estimates for the Eldis^2 and Dispersion scores. C-score reliabilities were computed as standardized squared distances (cf. Cohen, Cohen, West & Aiken, 2003, equation 2.2.4).

While the C-scores were least reliable (r_{xx} ranged from .67 to .82) and the Elevation based scores were most reliable (r_{xx} ranged from .97 to .98), the D^2 scores were also highly reliable (r_{xx} ranged from .93 to .96). The lower reliabilities of the C-scores are consistent with the computation of C-scores as the differences of standardized scores (i.e., C-scores computed as the difference of difference scores), while the Elevation, Eldis^2, and Dispersion scores are computationally simpler. The result that the D^2-scores were more reliable than the C-scores, yet less reliable than the elevation-based metrics, is consistent with the mathematical derivations showing that D^2 scores represent a function of the refined scores (C, Eldis^2, Dispersion scores). To presage remaining analyses, it may be counterproductive to assume that D^2-scores are superior to C-scores simply because of the greater reliability of D^2-scores.

Profile Score Correlations. The last four columns of Table 3 report correlations among the profile similarity metrics computed for each of the LKT scales. Correlations involving the D^2 and Eldis^2 scores were reflected so that superior performance on these metrics would correlate positively with other indices of LKT performance and experience (i.e. rank). These results are consistent with expectations based on our analysis of Equations 2 and 3 in showing substantial correlations among D^2, Eldis^2, Dispersion and C-scores.

Because D^2 is a function of the C-score, Dispersion and Eldis^2 metrics, we regressed D^2 on these metrics for each scale. Table 4 summarizes the results. For both LKT scales, virtually all the variance in D^2 is accounted for by the linear combination of these variables, $R^2 > .98$; the residual variance reflects the complicated interactions shown in Equation 3. The standardized weights reported in Table 4 estimate the extent to which D^2 variance is associated with each metric.

From a variance perspective, the results in Tables 3 and 4 show that LKT D^2 scores (when reflected so that superior performance is associated with higher values):

- Primarily reflected variance of the Eldis^2 scores, ($r > .90$; $\beta > .79$);
- Largely reflected variance of the C-scores, (r ranged from.43 to.82; β ranged from .27 to .45); and
- Moderately reflected the variance of the Dispersion scores, (r ranged up to .43; β ranged from -.26 to -.29).

In addition, the results reported in Table 3 show that the D^2 and Elevation scores were highly correlated, (r ranged from .63 to .70). Consistent with the formulaic analysis, these results show that the use of distance measures confounds variance that is associated with individual differences (i.e., variance) in the elevation, shape and dispersion of respondent rating profiles. It follows that distance metrics are preferable only if they represent the optimal combination of the C-scores, Eldis^2 and Dispersion scores.

Table 3. LKT Scale Descriptives and Correlations.

Scores	r_{xx}	*M*	*SD*	Profile Correlations[a]			
				Eldis^2	Elevation	Dispersion	D^2
LKT Trait Scale (Expert Key)							
C-score	.82	0.53	0.32	.60**	.52**	.42**	.82**
Eldis^2	.97	3.24	5.57		.71**	.61**	.91**
Elevation	.97	5.58	1.78			.35**	.70**
Dispersion	.86	2.10	0.74				.43**
D^2	.93	7.29	6.70				
LKT Skill Scale (Expert Key)							
C-score	.70	0.39	0.32	.40**	.26**	.27**	.53**
Eldis^2	.98	3.17	5.60		.67**	.32**	.94**
Elevation	.98	6.08	1.77			-.10	.70**
Dispersion	.86	1.67	0.71				.08
D^2	.95	6.34	6.02				
LKT Trait Scale (Consensus Key)							
C-score	.80	0.55	0.31	.59**	.56**	.44**	.73**
Eldis^2	.96	3.16	4.95		.62**	.63**	.90**
Elevation	.97	5.58	1.78			.35**	.63**
Dispersion	.86	2.10	0.74				.34**
D^2	.95	6.44	5.13				
LKT Skill Scale (Consensus Key)							
C-score	.67	0.41	0.30	.39**	.27**	.28**	.43**
Eldis^2	.98	3.13	5.17		.60**	.36**	.92**
Elevation	.98	6.08	1.77			-.10	.65**
Dispersion	.86	1.67	0.71				.02
D^2	.96	5.79	5.20				

[a] D^2 and Eldis^2 reflected to compute correlations. *N* ranged from 306 to 328. ** $p < .001$.

Table 4. D^2 Regressed on C-score, $Eldis^2$ and Dispersion indices.

Scale	R^2	*F*	*df*	Sig	β-weights		
					C-score	$Eldis^2$	Dispersion
Trait	.98	5213.7	3/284	.000	.45	.79	-.26
Skills	.99	6398.4	3/271	.000	.27	.90	-.29

Note area: All Beta coefficients significant at $p < .001$.

Profile Score Validities. Table 5 reports correlations between the profile similarity metrics computed for each of the LKT scales and the rank and personality criteria. The results indicated that the C-scores had substantially higher correlations with rank than any of the other profile similarity metrics. The C-scores also had high correlations with Openness, Conscientiousness, and Agreeableness. In addition, the Elevation, $Eldis^2$, D^2-scores had substantial correlations with Openness, Conscientiousness and Agreeableness. Although we had not hypothesized relationships, the highest correlations with the personality metrics were associated with the elevation scores.

Table 5. LKT Scale Validities for Scores Based on Expert Key.

Scores	Rank	Mini-markers				
		Open	Consc	Extrv	Agree	Neur-R
			LKT Trait Scale			
C-score	$.55^{***}$	$.40^{***}$	$.37^{***}$	.13	$.39^{***}$	.10
$Eldis^2$ (reflected)	$.25^{***}$	$.29^{***}$	$.32^{***}$	.11	$.30^{***}$	-.04
Elevation	$.19^{***}$	$.45^{***}$	$.40^{***}$	.12	$.37^{***}$	.01
Scatter	$.13^{*}$	$.15^{*}$	$.17^{**}$	$.15^{*}$	$.17^{*}$	.01
D^2 (reflected)	$.41^{***}$	$.38^{***}$	$.38^{***}$	.11	$.37^{***}$	.00
			LKT Skill Scale			
C-score	$.67^{***}$	$.21^{**}$	$.28^{***}$	.01	$.22^{***}$	.11
$Eldis^2$ (reflected)	$.23^{***}$	$.23^{***}$	$.28^{***}$	.00	$.23^{***}$	-.05
Elevation	.11	$.41^{***}$	$.39^{***}$	.09	$.31^{***}$	.01
Scatter	.11	$-.16^{*}$	-.10	-.06	$-.14^{*}$	-.09
D^2 (reflected)	$.36^{***}$	$.32^{***}$	$.36^{***}$	.01	$.33^{***}$	-.01
			Rank			
Rank		.12	$.21^{**}$	.03	$.21^{**}$	.12

Note: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$. *n* ranged from 223 to 323

Table 6 reports validities that have been corrected for attenuation of reliability. Results show that the C-scores generally have higher correlations with both rank and personality (i.e., Openness, Conscientiousness and Agreeableness) than any of the other LKT metrics.

Table 6. LKT Scale Validities Corrected for Attenuation of Reliability.

Scores	Rank	Mini-markers				
		Open	Consc	Extrv	Agree	Neur-R
			LKT Trait Scale			
C-score	.61***	.44***	.41***	.14	.43***	.11
Eldis^2	.25***	.29***	.32***	.11	.30***	-.04
Elevation	.19***	.46***	.41***	.12	.38***	.01
Scatter	.14*	.16*	.18**	.16*	.18*	.01
D^2	.43***	.39***	.39***	.11	.38***	.00
			LKT Skill Scale			
C-score	.80***	.25***	.33***	.01	.26***	.13
Eldis^2	.23***	.23***	.28***	.00	.23***	-.05
Elevation	.11	.41***	.39***	.09	.31***	.01
Scatter	.12	-.17*	-.11	-.06	-.15*	-.10
D^2	.37***	.33***	.37***	.01	.33***	-.01

To determine if any of the supplemental component metrics accounted for incremental variance to the C-scores for predicting rank and personality, we conducted hierarchical regression analyses with the C-scores entered in the first step, followed by either, the Eldis^2, the Elevation, or the Dispersion metric in the second step. Because this method resulted in 24 regression analyses (24 = 2 scales x 3 LKT scale metrics in the 2nd step x 4 criteria), we also conducted 8 summary regression analyses (8 = 2 scales x 4 criteria) in which the C-scores were entered in the first step followed by the three supplemental variables in the second step.

Because essentially parallel results were obtained from the two approaches, we report results for only the 8 regression analyses in Table 7 in order to provide a comprehensible summary of the findings. In general, the results showed only modest effects for the supplemental component metrics in predicting the rank criterion, but quite substantial effects for Elevation in accounting for variance in the personality indices of Openness, Conscientiousness and Agreeableness.

Table 7. Criteria Regressed on LKT C-scores (Step 1) and supplemental variables (Step 2).

Criteria	R^2	ΔF	*df*	Sig		β-weights from Step 2		
	Step1/2	Step1/2	Step1/2	Step1/2	C-score	$Eldis^2$ (reflected)	Dispersion	Elevation
						LKT Trait Scores		
Rank	.31/.32	140/2.74	1/318, 1/315	.000/.043	.62***	.07	-.15*	-.08
Openness	.16/.24	43.5/7.68	1/232, 3/229	.000/.000	.26***	-.16	.01	.41***
Conscientiousness	.14/.20	37.0/5.57	1/232, 3/229	.000/.001	.23**	-.00	-.04	.30**
Agreeableness	.15/.19	41.5/3.55	1/232, 3/229	.000/.015	.28***	-.04	-.02	.26**
						LKT Skill Scores		
Rank	.46/.49	253/4.88	1/298, 3/295	.000/.002	.72***	.12	-.18***	-.12
Openness	.05/.22	10.8/14.85	1/212, 3/209	.001/.000	.16*	-.06	-.17*	.41***
Conscientiousness	.09/.20	19.6/10.3	1/212, 3/209	.000/.000	.23***	.06	.16*	.27**
Agreeableness	.06/.18	13.1/9.9	1/212, 3/209	.000/.000	.22**	.08	-.24***	.20

* $p < .05$; ** $p < .01$; *** $p < .001$.

LKT Discussion

Based on expectations that leaders gain tacit knowledge regarding the importance of leader-relevant skills and traits through experience and reflections upon those experiences (Polanyi, 1966; Wagner & Sternberg, 1985; Sternberg et al., 2000), we expected that LKT metrics would correlate with rank. We evaluated this hypothesis by correlating a variety of LKT metrics with rank. Very substantial correlations were observed between the LKT C-scores and rank, all observed $r > .55$ and all corrected $r > .61$. Therefore, our first hypothesis (H1) that LKT metrics would correlate with rank was verified.

Based on conceptualizations regarding leadership and personality (e.g., Yukl, 2002; Bartone, Snook, & Tremble, 2002; Bartone, Eid, Johnsen, Laberg, & Snook, 2009), we expected that LKT metrics would also correlate with personality traits. Support for this expectation was demonstrated by substantial correlations between three of the personality constructs (Openness, Conscientiousness and Agreeableness) and the LKT C-scores, Elevation scores and D^2-scores. Thus our second hypothesis (H2) that LKT metrics would correlate with leadership-oriented personality traits was verified.

In general, LKT validities were highest for the C-scores. Somewhat paradoxically, this pattern was observed despite the C-scores having the lowest reliabilities of any of the LKT indices. These results suggest that expertise is best assessed through the use of C-scores, or by computing standardized differences scores so that elevation and variance differences removed for the LKT. At least for the LKT, C-scores appear superior to D^2-scores based on their correlations with conceptually relevant variables (i.e., rank as an indicators of expertise, and personality traits associated with leadership expertise).

In addition, we obtained higher validities for the LKT Skill C-scores than the Trait C-scores. The higher validities for the Skill C-scores may suggest that knowledge regarding the frequency of skills required for Army leadership accrues as a result of exposure to Army leadership experiences. Possibly, knowledge of the importance of leadership traits reflects broader experiences and may be more useful for predicting performance as officers. This interpretation is also consistent with the observation of higher correlations between LKT Trait scores and personality related leadership traits than between the LKT Skill scores and those same personality factors.

While we did not have strong theoretical expectations for the substantial correlations between elevation scores and personality constructs (r up to .45), the results demonstrate that it may be necessary to utilize different LKT metrics for different criteria. In retrospect, it seems reasonable that more agreeable individuals would be more likely to endorse (i.e., agree) that a wide variety of traits and skills are needed for effective leadership. Similar expectations could be posited for individuals who are higher on openness and conscientiousness. However, it is also possible that the stronger elevation and personality correlations stem from modulus biases in individual Likert rating profiless; so that the remaining fundamental correlation between the rating modulus in the personality and LKT measures remains as the strongest feature. In other words, people who elevate their scores on personality scales may also elevate their scores on SJTs. This would then be simply described as a method bias, or a method correlation. Notice that this method bias (if it exists) would tend to reduce the correlations for the D^2-scores and

Dispersion scores (an elevation bias would also affect dispersion), but leave the C-scores relatively unbiased.

From the perspective of the LKT application, the results suggest that adjective checklists may be easily converted into judgment tests that are valid against conceptually relevant criteria, thereby providing an objective measure of leadership knowledge. Despite the LKT scales being preliminary, the validity estimates are similar to or exceed correlations reported for many SJTs against performance criteria (McDaniel et al., 2001). In addition, the scales require only several minutes to complete, and may reduce self-enhancement or self-deception biases because they objectively assess knowledge by asking for descriptions of general knowledge rather than personal attributes, and therefore may support personnel selection goals. Due to the limited administration requirements of these scales, they could easily be lengthened to improve their reliability and validity.

GENERAL DISCUSSION

The mathematical analyses show that while distance measures partially reflect the similarities in form between respondent rating profiles and the scoring keys, distance measures introduce elevation and dispersion effects to their detriment. It is also clear from the mathematical analysis that under at least some conditions, distance scores can be highly saturated by elevation and dispersion effects. Under these conditions, distance scores may function poorly as indices of SJT performance.

However, the mathematical analyses leave open questions regarding the extent to which distance scores and other measures of overall agreement are saturated by elevation and dispersion effects. Regarding this last point, the analyses conducted using LKT data are important because they showed very powerful effects in favor of C-scores over distance-based metrics. As such, understanding relationships among correlation, elevation and dispersion metrics is relevant to any SJT application that utilizes distance metrics to score respondent ratings data (i.e., rating profiles). Therefore, these analyses indicate that is generally advisable to use C-scores to score judgment tests that utilize respondent ratings.

Incidentally, if the correlations with rank had been lower for the LKT C-scores than for the distance metric, and especially if the C-score validities had been much lower, then interpretations regarding the LKT would have been dramatically altered. This is because high D^2-score validities, coupled with low C-score validities, could only occur if the D^2-score validities had reflected either elevation or dispersion effects. In this circumstance, it would be incorrect and misleading to assert that the LKT would be assessing knowledge as intended.

Moreover, this reasoning applies to most judgment tests that have been described in the literature. It follows from the mathematical derivations that judgment test scores that quantify absolute agreement with a scoring standard (e.g., distance metrics or endorsement ratios) can only be valid if the scores reflect meaningful differences in elevation, dispersion, or association (C-scores). However, measures of elevation and dispersion have not been used to score SJTs because elevation and dispersion effects have not been recognized as meaningful. It follows that judgment test distance scores that are valid only due to individual differences in elevation and dispersion effects, if there any, would not be acting as intended. Of course, it would be possible

to intentionally design such a scale, but to claim that a conventional judgment test is valid when the scores primarily assess elevation or dispersion effects would be highly misleading.

At this point it is still speculative, but it seems likely that the various component scores would be differentially affected by the collection of data under research and high-stakes conditions. This is likely because it is relatively easy for a respondent to vary the elevation and dispersion of their ratings, while it appears difficult to substantially vary the shape of one's ratings, other than to decrease its consistency (i.e., correlation) with the scoring standard. So elevation and dispersion effects, if strong, may be related to self-distortion affects (i.e., faking on traditional personality inventories).

As suggested above, we believe that few researchers would intentionally use SJT metrics that combine shape, elevation and dispersion effects in arbitrary ways. The statistical analyses using the LKT database indicate that at least for that database the extent of the interaction among the metrics was substantial and reduced the meaningful assessment of trait and skill knowledge. Additional analyses using other SJT databases would likely shed more light on the extent to which the use of uncontrolled mixtures of distance metrics reduces the validity of SJTs.

Although our mathematical analysis cannot estimate the extent to which distance measures are confounded by elevation and dispersion effects in typical judgment tests, results based on available data clearly indicate that correlation scores are superior to distance scores for some purposes. Given that C-scores are optimized measures of least squares distance scores, the mathematical underpinnings alone might suggest that they have greater validity. By empirically verifying this expectation, the present report suggests that using profile measures (e.g., C-scores) as opposed to unstandardized distance metrics should optimize SJT validity estimates for many applications.

REFERENCES

Bartone, P., Eid, J., Johnsen, B., Laberg, J., & Snook, S. (2009). Big five personality factors, hardiness, and social judgment as predictors of leader performance. *Leadership & Organization Development Journal, 30*(6), 498-521.

Bartone, P., Snook, S., & Tremble, T. (2002). Cognitive and personality predictors of leader performance in West Point cadets. *Military Psychology, 14*(4), 321-338.

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The Nature of Expertise.* Hillsdale, NJ: Lawrence Earlbaum Associates Inc.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50,* 456-473.

Cullen, M. J., Sackett, P. R., & Lievens, F. (2006). Threats to the operational use of situational judgment tests in the college admission process. *International Journal of Selection and Assessment, 14,* 142-155.

Grim, A. (2010). Use of Situational Judgment Test to Measure Individual Adaptability in Applied Settings. Unpublished Mater's Thesis. George Mason University.

Legree, P J. (1995). Evidence for an oblique social intelligence factor established with a Likert based testing procedure. *Intelligence, 21,* 247-266.

Legree, P. J., Psotka J., Bludau, T. M. & Gray, D. M. (2008, April). Assessing Occupational Knowledge using SJTs derived from Job Analysis Questionnaires. Paper presented at the Society for Industrial and Organizational Psychology (SIOP) 23rd Annual Conference, San Francisco, CA.

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. (2005). Using consensus based measurement to assess emotional intelligence. In R. Schulze & R. D. Roberts, *Emotional Intelligence: An International Handbook* (pp 155-180). Berlin, Germany: Hogrefe & Huber.

Mayer, J. D., Caruso D. R. & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence,* 27, 267-298.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braveman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86,* 730-740.

McDaniel, M.A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103-113.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640-647.

Pleskac, T., & Busemeyer, J. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review, 117,* 864-901.

Polanyi, M. (1966). *The Tacit Dimension*. New York: Doubleday.

Sternberg, R.J., Forsythe, G.B., Hedlund, J., Horvath, J. A., Wagner, R.K., Williams, W.M., Snook, S., & Grigorenko, E.L. (2000). *Practical intelligence in everyday life*. New York: Cambridge University Press.

Stevens, S. S. (1975) *Psychophysics: Introduction to its perceptual, neural and social prospects.* Oxford, England: John Wiley & Sons.

Yukl, G. (2002). *Leadership in Organizations.* 5th Edition. Englewood Cliffs, NJ: Prentice Hall.

Wagner**,** R. K., & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49*, 436-458.

Weis, S (2008). *Theory and Measurement of Social Intelligence as a Cognitive Performance Construct.* (Doctoral Thesis, Otto-von-Guericke University, Magdeburg, Germany). Retrieved from http://diglib.uni-magdeburg.de/Dissertationen/2008/susweis.pdf

Appendix A

Relationships among D^2, C, sd_x and $\Delta_{elevation}$ as individual difference metrics for Judgment Tests constructed to provide rating data.

Definitions: Let vector **X** correspond to n observed ratings provided by respondent X, and vector **K** correspond to n ratings used to score **X**. The following derivations link D^2 to $\Delta_{elevation}$, sd_x and C-scores.

$D^2 = \sum(X_i - K_i)^2/n$ for item i = 1 to n	Conventional distance formula (Equation 1 in text)
$= \sum((x_i + X_{mean}) - (k_i + K_{mean}))^2/n$	Substitutions center X and K: $x_i = X_i - X_{mean}$ thus $X_i = x_i + X_{mean}$; $k_i = K_i - K_{mean}$ thus $K_i = k_i + K_{mean}$
$= \sum(x_i + X_{mean} - k_i - K_{mean})^2/n$	Distributive Property
$= \sum(x_i - k_i + (X_{mean} - K_{mean}))^2/n$	Rearrange and group
$= \sum(x_i - k_i + \Delta)^2/n$	Substituting Δ for $X_{mean} - K_{mean}$
$= 1/n\sum(x_i - k_i + \Delta)^2$	Constant multiplication property of sums
$= 1/n\sum(x_i^2 + k_i^2 + \Delta^2 - 2x_ik_i + 2x_i\Delta - 2k_i\Delta)$	Binomial expansion
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i + \sum 2x_i\Delta - \sum 2k_i\Delta)$	Expansion property of sums
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i + 2\Delta\sum x_i - 2\Delta\sum k_i)$	Constant multiplication property of sums
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i + 2\Delta 0 - 2\Delta 0)$	$\sum x_i = 0$ & $\sum k_i = 0$ because x & k are centered
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i + 0 - 0)$	Multiplication property of zero
$= 1/n(\sum x_i^2 + \sum k_i^2 + \sum\Delta^2 - \sum 2x_ik_i)$	Additive property of zero
$= 1/n\sum\Delta^2 + 1/n(\sum x_i^2 + \sum k_i^2 - \sum 2x_ik_i)$	Regrouping property of sums
$= \Delta^2 + 1/n\sum(x_i^2 + k_i^2 - 2x_ik_i)$	Summation of a constant: Substitutes $1/n\sum\Delta^2 = 1/n(n\Delta^2) = \Delta^2$
$= \Delta^2 + 1/n\sum(x_i - k_i)^2$	Provides binomial solution (Equation 2 in text)
$= \Delta^2 + 1/n\sum(x_i^2 + k_i^2 - 2x_ik_i)$	From two steps above
$= \Delta^2 + 1/n(\sum x_i^2 + \sum k_i^2 - \sum 2x_ik_i)$	Expansion property of sums

$= \Delta^2 + 1/n(\sum x_i^2 + \sum k_i^2 - 2\sum x_i k_i)$	Constant multiplication property of sums
$= \Delta^2 + 1/n(sd_x^2(n-1) + sd_k^2(n-1) - 2\sum x_i k_i)$	Substitutions based on statistical formulas re variance: $\sum x_i^2 = sd_x^2(n\text{-}1)$ & $\sum k_i^2 = sd_k^2(n\text{-}1)$
$= \Delta^2 + 1/n(sd_x^2(n-1) + sd_k^2(n-1) - 2\sum z_{xi} sd_x z_{ki} sd_k)$	Substitutions based on statistical formulas re z-scores: $x_i = z_{xi} sd_{xi}$ & $k_i = z_{ki} sd_k$
$= \Delta^2 + 1/n(sd_x^2(n-1) + sd_k^2(n-1) - 2sd_x sd_k \sum z_{xi} z_{ki})$	Constant multiplication property of sums
$= \Delta^2 + 1/n(sd_x^2(n-1) + sd_k^2(n-1) - 2sd_x sd_k r(n-1))$	Substitutions based on formulas re the product moment correlation: $r = \sum z_{xi} z_{ki}/(n - 1)$ thus $\sum z_{xi} z_{ki} = r(n - 1)$
$= \Delta^2 + (n-1)/n(sd_x^2 + sd_k^2 - 2sd_x sd_k r)$	Rearrangement of terms
$= \Delta^2 + ((n-1)/n(sd_x^2 + sd_k^2 - 2sd_x sd_k C))$	Designates C as representing correlation scores based on r (Provides Equation 3)
$= (X_{mean} - K_{mean})^2 + ((n-1)(sd_x^2 + sd_k^2 - 2sd_x sd_k C))/n$	Alternate form of Equation 3

In equation 3, sd_k is constant for applications and C is an individual difference measure of maximal performance. Because equation 3 has a binomial form, manipulations of sd_x will have a non-monotonic impact on D^2. Differentiating the function D^2 with respect to sd_x provides:

$d(D^2)/d(sd_x) = ((n-1)/n(2sd_x - 2sd_k C))$

Solving for the inflection point:

$0 = ((n-1)/n(2sd_x - 2sd_k C))$	Formula for inflection point
$0 = 2sd_x - 2sd_k C$	Dividing constants
$0 = sd_x - sd_k C$	Dividing constants
$sd_x = sd_k C$	Provides value of sd_x that minims D^2

Appendix B

Relationships between the variance of items in scoring key **K** and the variance in individual expert rating profiles used to develop the scoring profile, **K**.

Note: **K** is derived by averaging ratings for n items obtained from p experts. The rating profile obtained for each expert *j* is designated as $\mathbf{E_j}$, and value $E_{i,j}$ designates response for item i by expert j.

For each expert j:

$E_{mean\_j} = \sum(E_{i,j}/n)$ for i = 1 to n items	Conventional definition.	B1

For each value K_i:

$K_i = \sum E_{i,,j}/p$ for j = 1 to p experts, i is constant	As defined above.	B2
$= 1/p\sum E_{i,,j}$	Constant multiplication property of sums	B3
$= 1/p\sum(e_{i,j} + E_{mean\_j})$	Substitutes centered vales for $E_{i,j}$, i.e., $e_{i,j} = E_{i,j} - E_{mean\_j}$ thus $E_{i,j} = e_{i,j} + E_{mean\_j}$	B4

In addition, K_{mean} equals the mean item value in *K* as well as the mean expert mean. This is shown by:

$K_{mean} = \sum K_i/n$ for i = 1 to n items	Conventional definition	B5
$= \sum\sum(E_{i,j}/pn)$ for i = 1 to n items & j = 1 to p experts	Substitutes the above equality (B3): $K_i = E_{i,j}/p$	B6
$= 1/p\sum\sum(E_{i,j}/n)$ for i = 1 to n items & j = 1 to p experts	Constant multiplication property for sums	B7
$= 1/p\sum E_{mean\_j}$ for j = 1 to p experts	Substitutes above formula (B1): $E_{mean\_j} = \sum(E_{i,j}/n)$	B8

The following shows that k_i (centered K_i) equals the mean centered expert rating:

$k_i = K_i - K_{mean}$	Defines k_i as centered K_i	B9
$= 1/p\sum(e_{i,j} + E_{mean\_j}) - K_{mean}$ for j = 1 to p experts	Substitutes above equality	B10
$= 1/p\sum e_{i,j} + 1/p\sum E_{mean\_j} - K_{mean}$	Regrouping property of sums	B11
$= 1/p\sum e_{i,j} + 1/p\sum E_{mean\_j} - 1/p\sum E_{mean\_j}$	Substitute above equality (B8)	B12
$= 1/p\sum e_{i,j}$	Terms cancel	B13

Then the square of each value, k_i, is given by:

$k_i^2 = k_i k_i$	Definition	B14
$= 1/p\sum e_{i,j} 1/p\sum e_{i,h}$ for j & h = 1 to p	Substitute above equality (B13)	B15
$= 1/p^2\sum e_{i,j}\sum e_{i,h}$ for j & h = 1 to p	Rearrange constant in summation terms	B16
$= 1/p^2\sum\sum e_{i,j}e_{i,h}$ for j & h = 1 to p	Rearrange summation terms	B17

And the variance of items in **K,** sd^2_k, is given by:

$sd^2_k = 1/(n\text{-}1)\sum k_i^2$ for i = 1 to n items	Formula for the variance of k	B18
$= (1/(n–1))\sum(1/p^2)\sum\sum(e_{i,j}e_{i,h}))$ for i = 1 to n, and j & h = 1 to p	Substitutes above equality (B17)	B19
$= (1/(p^2(n–1)))\sum\sum\sum(e_{i,j}e_{i,h})$ for i = 1 to n, and j & h = 1 to p	Constant multiplication property for sums	B20
$= (1/(p^2(n–1)))\sum\sum\sum(sd_{i,j}z_{i,j}sd_{i,h}z_{i,h}))$ for i = 1 to n and j & h = 1 to p	Statistical substitution: $e_{i,j} = sd_{i,j}z_{i,j}$; and $e_{i,h} = sd_{i,h}z_{i,h}$	B21
$= (1/(p^2(n–1)))\sum\sum\sum(sd_{i,j}sd_{i,h}z_{i,j}z_{i,h})$ for i = 1 to n and j & h = 1 to p	Rearrange terms	B22

Equation B22 shows that the *sd* terms weight the z-scores corresponding to each expert. Equally weighting all experts by setting all $sd_{i,j} = sd_c$ (constant) provides:

$sd^2_k = (1/((n\text{-}1)p^2))\sum\sum\sum(sd_c sd_c z_{ij} z_{ih})$ for i = 1 to n; and j & h = 1 to p	Equally weights sd for all experts.	B23
$= sd_c sd_c(1/((n\text{-}1)p^2))\sum\sum\sum(z_{ij} z_{ih})$ for i = 1 to n; and j & h = 1 to p	Constant multiplication property of sums	B24
$= (sd_c^2/((n\text{-}1)p^2))\sum\sum\sum(z_{ij} z_{ih})$ for i = 1 to n; and j & h = 1 to p	Combine constant terms	B25
$= (sd_c^2/((n\text{-}1)p^2))\sum\sum\sum(z_{ij} z_{ih})$ for j & h = 1 to p and i = 1 to n;	Exchange order of double sums	B26
$= (sd_c^2/((n\text{-}1)p^2))\sum\sum(r_{j,h}(n\text{-}1))$ for j & h = 1 to p	Substitutes statistical equality: $\sum(z_{ij} z_{ih}) = r_{j,h}(n\text{-}1)$ for i = 1 to n	B27
$= (sd_c^2(n\text{-}1)/((n\text{-}1))p^2)\sum\sum r_{j,h}$ for j & k = 1 to p	Constant multiplication property of sums	B28
$= (sd_c^2/p^2)\sum\sum r_{j,h}$ for j & h = 1 to p	Terms cancel	B29

And

$$sd_k = (sd_c/p)(\sum\sum r_{j,h})^{1/2} \text{ for } j \text{ \& } h = 1 \text{ to } p$$

Note that $\sum\sum r_{j,h}$ is the p by p correlation matrix of expert ratings with 1's in the diagonal. Thus the sum of its entries must be less or equal to p^2, and will only equal p^2 if all the expert ratings are perfectly correlated. Rearranging the formula provides Equation 5:

$$sd_k = (sd_c/p)(1 + \text{sqrt}(r_{mean})(p\text{-}1))$$

where r_{mean} is the mean intercorrelation of expert rating profiles.

For moderate values of r_{mean} , this function will quickly approach its asymptote, $(r_{mean})^{1/2}$. For example, if r_{mean} = .64, then the asymptote is .80. For following table estimates the ratio of sd_k to sd_c for various numbers of experts, *p*. These computations show that sd_k will be usually much smaller than the *sd*'s of the rating profiles of individual experts.

p	Ratio of sd_k to sd_c
2	.90
5	.84
10	.82
20	.81

$p = 2$, $sd_k = (sd_c/2)(1 + sqrt(.64)(1)) = .90sd_c$

$p = 5$, $sd_k = (sd_c/5)(1 + sqrt(.64)(4)) = .84sd_c$

$p = 10$, $sd_k = (sd_c/10)(1 + sqrt(.64)(9)) = .82sd_c$

$p = 20$, $sd_k = (sd_c/20)(1 + sqrt(.64)(19)) = .81sd_c$